Information Systems 38 (2013) 1070–1083

Contents lists available at SciVerse ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/infosys

SVOIS: Support Vector Oriented Instance Selection for text classification

Chih-Fong Tsai*, Che-Wei Chang

Department of Information Management, National Central University, Taiwan

ARTICLE INFO

Article history:
Received 18 June 2012
Accepted 10 May 2013
Recommended by F. Carino Jr.
Available online 22 May 2013

Keywords:
Instance selection
Data reduction
Text classification
Machine learning
Support vector machines

ABSTRACT

Automatic text classification is usually based on models constructed through learning from training examples. However, as the size of text document repositories grows rapidly, the storage requirements and computational cost of model learning is becoming ever higher. Instance selection is one solution to overcoming this limitation. The aim is to reduce the amount of data by filtering out noisy data from a given training dataset. A number of instance selection algorithms have been proposed in the literature, such as ENN, IB3, ICF, and DROP3. However, all of these methods have been developed for the *k*-nearest neighbor (*k*-NN) classifier. In addition, their performance has not been examined over the text classification domain where the dimensionality of the dataset is usually very high. The support vector machines (SVM) are core text classification techniques. In this study, a novel instance selection method, called Support Vector Oriented Instance Selection (SVOIS), is proposed. First of all, a regression plane in the original feature space is identified by utilizing a threshold distance between the given training instances and their class centers. Then, another threshold distance, between the identified data (forming the regression plane) and the regression plane, is used to decide on the support vectors for the selected instances. The experimental results based on the TechTC-100 dataset show the superior performance of SVOIS over other state-of-the-art algorithms. In particular, using SVOIS to select text documents allows the *k*-NN and SVM classifiers perform better than without instance selection.

1. Introduction

The number and size of online information collections are increasing rapidly, meaning that text classification (or categorization) has become one of the major techniques for managing large scale text repositories. The aim of text classification is to automatically classify documents into a fixed set of pre-defined categories, in which text documents are first processed by the natural language processing technique, and then each document is represented by a *d*-dimensional feature vector. Next, specific classification techniques, such as support vector machines, can be used for model learning and classifying of text documents [18,20,27,31].

Data pre-processing is one of the most critical steps in data mining and knowledge discovery in database (KDD) techniques, performed to ensure good quality data mining. Feature selection (or dimensionality reduction) has been extensively studied in the text classification literature; for examples see [10,32]. This is because the dimensionality of the extracted textual features that represent text documents (i.e., *tf-idf*) is usually very large, say 10,000.

However, the size of today's text collections often exceeds the size of the datasets which the current software and/or hardware handle properly. In spite of this, there have been few studies focused on instance selection (or data reduction) for text classification. That is, if too many instances (i.e., documents) are adopted, it can result in large memory requirements and slow execution speed,

* Corresponding author. Tel.: +886 3 422 7151; fax: +886 3 4254604.
E-mail address: cftsai@mgt.ncu.edu.tw (C.-F. Tsai).

0306-4379/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.is.2013.05.001

and can cause over-sensitivity to noise [21,30]. Furthermore, one problem with using the original data points is that there may not be any located at the precise points that would make for the most accurate and concise concept description [23].

In addition to feature selection, instance selection (or data reduction) is another important data pre-processing step in the KDD process. The aim of instance selection is to reduce the data size by filtering out noisy data from a given dataset, which would otherwise increase the likelihood of degrading the mining performance. In particular, instance selection is used to shrink the amount of data, so data mining algorithms can be applied to the reduced dataset. Sufficient results can be achieved if the selection strategy is appropriate [26].

Outlier detection involves the finding of observations that lie an abnormal distance from other values in a random sample from a given population. Outliers have traditionally been defined as unusual observations (or bad data points) that are far removed from the mass of data [1,3]. Consequently, classifiers trained by selected instances as a subset of original instances can provide relatively good performances. Outlier detection is also a critical KDD task [19], and filtering out the detected outliers is very useful for obtaining good mining results. From the data mining perspective, the aim of instance selection is similar to that of outlier detection [22].

In this study, a novel instance selection method, called Support Vector Oriented Instance Selection (SVOIS) is proposed for text classification. SVOIS mainly borrows the idea of support vector machines (SVM). The support vectors in SVM are used for binary classification decisions [28]. Specifically, given a training dataset, each training vector is associated with one of two different classes. During the training stage, the input vectors are mapped into a new higher dimensional feature space. Then, an optimal separating hyperplane is constructed in the new feature space. All vectors lying on one side of the hyperplane can be regarded as class 1, and all vectors lying on another side are class 2. The training instances that lie closest to the hyperplane in the transformed space are called support vectors. The number of these support vectors is usually small compared to the size of the training set and they determine the margin of the hyperplane, and thus the decision surface. In order to produce good generalizations, the SVM maximizes the margin of the hyperplane and diminishes the number of support vectors for it.

However, the major limitations of the SVM are speed and size, both in training and testing [6,7]. In other words, the computational cost necessary to identify the hyperplane and support vectors in a new and very high dimensional feature space is excessive.

Unlike the SVM, the SVOIS attempts to find the support vectors in the original feature space through a linear regression plane, where the instances to be selected as the support vectors need to satisfy two criteria. The first one is that the distances between the original instances and their class centers need to be smaller than a predefined value. Then, the instances fulfilling this criterion are regarded as the regression data in order to identify a regression plane. The second criterion is based on the distances between the regression data and the regression plane, which is like the margin in the SVM. These distances need to be larger than a pre-defined value. The regression data fulfilling this criterion are called support vectors and are used for classifier training and classification. Specifically, these two types of distances should be neither too long, so that all instances are selected, nor too short, leading to very few support vectors (cf. Section 3).

To the best of our knowledge, SVOIS is the first method to select redundant instances (i.e., documents) for text classification. Current state-of-the-art instance selection techniques (cf. Section 2) are only assessed for classification performance over the datasets containing very low dimensionality. They perform poorly in the context of text classification. In contrast, using SVOIS to filter out unimportant documents from the given training dataset allows two well-known classifiers (i.e., SVM and k-NN) to perform better than the ones without instance selection and the same ones followed by state-of-the-art instance selection techniques.

The rest of this paper is organized as follows. Section 2 provides an overview of related literatures, including the concept of instance selection and four well-known instance selection algorithms, which are ENN, IB3, DROP3, and ICF. The proposed SVOIS method for text classification is introduced in Section 3. In Section 4, the experimental results based on a public text classification dataset are present. Finally, some conclusions are offered in Section 5.

2. Instance selection

Instance selection can be defined as follows. Given a dataset D composed of a training set T and testing set U, let X_i be the ith instance in D, where $X_i = (X_1, X_2, \ldots, X_m)$ which contains m different features. Let $S \subset T$ be the subset of selected instances resulting from the execution of an instance selection algorithm. Then, U is used to test a classification technique trained by S [8,12].

There are a number of studies in the literature related to proposing instance selection methods to obtain better mining quality. Specifically, Pradhan and Wu [24] and Jankowski and Grochowski [17] surveyed several relevant selection techniques, which can be divided into three application-type groups. These include noise filters, condensation algorithms, and prototype searching algorithms. On the other hand, extensive comparative experiments have been conducted by Wilson and Martinez [30] and Brighton and Mellish [5]. They utilized Iterative Case Filtering (ICF) and Decremental Reduction Optimization Procedure 3 (DROP3) and cutting-edge instance selection algorithms, which allowed the k-NN classifier to perform better than with many other instance selection methods. Four well-known instance selection algorithms, ENN, IB3, ICF, and DROP3, are reviewed below.

2.1. ENN

The Edited Nearest Neighbor (ENN) approach [29] is a representative noise-filtering algorithm, in which S starts out the same as T, after which each instance in S is